The Role of Normalization in the Belief Propagation Algorithm[*]

Victorin Martin
INRIA Paris-Rocquencourt

Jean-Marc Lasgouttes
INRIA Paris-Rocquencourt

Cyril Furtlehner
INRIA Saclay

January 24, 2011

Abstract

An important part of problems in statistical physics and computer science can be expressed as the computation of marginal probabilities over a Markov Random Field. The belief propagation algorithm, which is an exact procedure to compute these marginals when the underlying graph is a tree, has gained its popularity as an efficient way to approximate them in the more general case. In this paper, we focus on an aspect of the algorithm that did not get that much attention in the literature, which is the effect of the normalization of the messages. We show in particular that, for a large class of normalization strategies, it is possible to focus only on belief convergence. Following this, we express the necessary and sufficient conditions for local stability of a fixed point in terms of the graph structure and the beliefs values at the fixed point. We also explicit some connexion between the normalization constants and the underlying Bethe Free Energy.

1 Introduction

We are interested in this article in a random Markov field on a finite graph with local interactions, on which we want to compute marginal probabilities. The structure of the underlying model is described by a set of discrete variables $\mathbf{x} = \{x_i, i \in \mathbb{V}\} \in \{1, \dots, q\}^{\mathbb{V}}$, where the set $\mathbb{V}$ of variables is linked together by so-called "factors" which are subsets $a \subset \mathbb{V}$ of variables. If $\mathbb{F}$ is this set of factors, we consider the set of probability measures of the form

$$p(\mathbf{x}) = \prod_{i\in\mathbb{V}} \phi_i(x_i) \prod_{a\in\mathbb{F}} \psi_a(\mathbf{x}_a), \tag{1.1}$$

[*] This work was supported by the French National Research Agency (ANR) grant Travesti (ANR-08-SYSC-017)

where $\mathbf{x}_a = \{x_i, i \in a\}$.

$\mathbb{F}$ together with $\mathbb{V}$ define the factor graph $\mathcal{G}$ (Kschischang et al., 2001), that is an undirected bipartite graph, which will be assumed to be connected. We will also assume that the functions ψ_a are never equal to zero, which is to say that the Markov random field exhibits no deterministic behavior. The set $\mathbb{E}$ of edges contains all the couples $(a, i) \in \mathbb{F} \times \mathbb{V}$ such that $i \in a$. We denote d_a (resp. d_i) the degree of the factor node a (resp. of the variable node i), and C the number of independent cycles of $\mathcal{G}$.

Exact procedures to compute marginal probabilities of p generally face an exponential complexity problem and one has to resort to approximate procedures. The Bethe approximation, which is used in statistical physics, consists in minimizing an approximate version of the variational free energy associated to (1.1). In computer science, the belief propagation (BP) algorithm (Pearl, 1988) is a message passing procedure that allows to compute efficiently exact marginal probabilities when the underlying graph is a tree. When the graph has cycles, it is still possible to apply the procedure, which converges with a rather good accuracy on sufficiently sparse graphs. However, there may be several fixed points, either stable or unstable. It has been shown that these fixed points coincide with stationary points of the Bethe free energy (Yedidia et al., 2005). In addition (Heskes, 2003; Watanabe and Fukumizu, 2009), stable fixed points of BP are local minima of the Bethe free energy. We will come back to this variational point of view of the BP algorithm in Section 6.

We discuss in this paper an aspect of the algorithm that did not get that much attention in the literature, which is the effect of the normalization of the messages on the behavior of the algorithm. Indeed, the justification for normalization is generally that it “improves convergence”. Moreover, different authors use different schemes, without really explaining what are the difference between these definitions.

The paper is organized as follows: the BP algorithm and its various normalization strategies are defined in Section 2. Section 3 deals with the effect of different types of messages normalization on the existence of fixed points. Section 4 is dedicated to the dynamic of the algorithm in terms of beliefs and cases where convergence of messages is equivalent to convergence of beliefs; moreover, it is shown that normalization does not change belief dynamic. In Section 5, we show that normalization is required for convergence of the messages, and provide some sufficient conditions. Finally, in Section 6, we tackle the issue of normalization in the variational problem associated to Bethe approximation. New research directions are proposed in Section 7.

2 The belief propagation algorithm

The belief propagation algorithm (Pearl, 1988) is a message passing procedure, which output is a set of estimated marginal probabilities, the beliefs $b_a(\mathbf{x}_a)$ (including single nodes beliefs $b_i(x_i)$). The idea is to factor the marginal probability at a given site as a product of contributions coming from neighboring factor nodes, which are the messages. With definition (1.1) of the joint probability measure, the updates rules read:

$$m_{a\to i}(x_i) \leftarrow \sum_{\mathbf{x}_{a\setminus i}} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j), \tag{2.1}$$

$$n_{i\to a}(x_i) \overset{\text{def}}{=} \phi_i(x_i) \prod_{a'\ni i, a'\neq a} m_{a'\to i}(x_i), \tag{2.2}$$

where the notation $\sum_{\mathbf{x}_s}$ should be understood as summing all the variables x_i, $i \in s \subset \mathbb{V}$, from 1 to q. At any point of the algorithm, one can compute the current beliefs as

$$b_i(x_i) \overset{\text{def}}{=} \frac{1}{Z_i(m)} \phi_i(x_i) \prod_{a\ni i} m_{a\to i}(x_i), \tag{2.3}$$

$$b_a(\mathbf{x}_a) \overset{\text{def}}{=} \frac{1}{Z_a(m)} \psi_a(\mathbf{x}_a) \prod_{i\in a} n_{i\to a}(x_i), \tag{2.4}$$

where $Z_i(m)$ and $Z_a(m)$ are the normalization constants that ensure that

$$\sum_{x_i} b_i(x_i) = 1, \qquad \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) = 1. \tag{2.5}$$

These constants reduce to 1 when $\mathcal{G}$ is a tree.

In practice, the messages are often normalized so that

$$\sum_{x_i=1}^{q} m_{a\to i}(x_i) = 1. \tag{2.6}$$

However, the possibilities of normalization are not limited to this setting. Consider the mapping

$$\Theta_{ai,x_i}(m) \overset{\text{def}}{=} \sum_{\mathbf{x}_{a\setminus i}} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} \Bigl[\phi_j(x_j) \prod_{a'\ni j, a'\neq a} m_{a'\to j}(x_j)\Bigr]. \tag{2.7}$$

A normalized version of BP is defined by the update rule

$$\tilde{m}_{a\to i}(x_i) \leftarrow \frac{\Theta_{ai,x_i}(\tilde{m})}{Z_{ai}(\tilde{m})}. \tag{2.8}$$

where $Z_{ai}(\tilde{m})$ is a constant that depends on the messages and which, in the case of (2.6), reads

$$Z_{ai}^{\text{mess}}(\tilde{m}) \overset{\text{def}}{=} \sum_{x=1}^{q} \Theta_{ai,x}(\tilde{m}). \tag{2.9}$$

In the remaining of this paper, (2.1,2.2) will be referred to as "plain BP" algorithm, to differentiate it from the "normalized BP" of (2.8).

Following Wainwright (2002), it is worth noting that the plain message update scheme can be rewritten as

$$m_{a\to i}(x_i) \leftarrow \frac{Z_a(m) b_{i|a}(x_i)}{Z_i(m) b_i(x_i)} m_{a\to i}(x_i), \tag{2.10}$$

where we use the convenient shorthand notation

$$b_{i|a}(x_i) \overset{\text{def}}{=} \sum_{\mathbf{x}_{a\setminus i}} b_a(\mathbf{x}_a).$$

This suggests a different type of normalization, used in particular by Heskes (2003), namely

$$Z_{ai}^{\text{bel}}(\tilde{m}) = \frac{Z_a(\tilde{m})}{Z_i(\tilde{m})}, \tag{2.11}$$

which leads to the simple update rule

$$\tilde{m}_{a\to i}(x_i) \leftarrow \frac{b_{i|a}(x_i)}{b_i(x_i)} \tilde{m}_{a\to i}(x_i). \tag{2.12}$$

The following lemma recapitulates some properties shared by all normalization strategies at a fixed point.

Lemma 2.1. *Let $\tilde{m}$ be such that*

$$\tilde{m}_{a\to i}(x_i) = \frac{\Theta_{ai,x_i}(\tilde{m})}{Z_{ai}(\tilde{m})}.$$

The associated normalization constants satisfy

$$Z_{ai}(\tilde{m}) = \frac{Z_a(\tilde{m})}{Z_i(\tilde{m})}, \qquad \forall ai \in \mathbb{E}, \tag{2.13}$$

and the following compatibility condition holds.

$$\sum_{\mathbf{x}_{a\setminus i}} b_a(\mathbf{x}_a) = b_i(x_i). \tag{2.14}$$

In particular, when $Z_{ai} \equiv 1$ (no normalization), all the Z_a and Z_i are equal to some common constant Z.

Proof. The normalized update rule (2.8), together with (2.3)–(2.4), imply

$$\sum_{\mathbf{x}_a \backslash x_i} b_a(\mathbf{x}_a) = \frac{Z_i Z_{ai}}{Z_a} b_i(x_i).$$

By definition of Z_a and Z_i, b_a and b_i are normalized to 1, so summing this relation w.r.t x_i gives (2.13) and the equation above reduces to (2.14). ∎

It is known (Yedidia et al., 2005) that the belief propagation algorithm is an iterative way of solving a variational problem, namely it minimizes over b the Bethe free energy $F(b)$ associated with (1.1).

$$F(b) \overset{\text{def}}{=} \sum_{a,\mathbf{x}_a} b_a(\mathbf{x}_a) \log \frac{b_a(\mathbf{x}_a)}{\psi_a(\mathbf{x}_a)} + \sum_{i,x_i} b_i(x_i) \log \frac{b_i(x_i)^{1-d_i}}{\phi_i(x_i)}. \tag{2.15}$$

Writing the Lagrangian of the minimization of (2.15) with b subject to the constraints (2.14) and (2.5), one obtains

$$\mathcal{L}(b,\lambda,\gamma) = F(b) + \sum_{\substack{i,a\ni i\\ x_i}} \lambda_{ai}(x_i)\Big(b_i(x_i) - \sum_{\mathbf{x}_a/x_i} b_a(\mathbf{x}_a)\Big) - \sum_i \gamma_i\Big(\sum_{x_i} b_i(x_i) - 1\Big).$$

The minima are stationary points of $\mathcal{L}(b,\lambda,\gamma)$ which correspond to

$$\begin{cases} b_a(\mathbf{x}_a) & = \dfrac{\psi_a(\mathbf{x}_a)}{e} \displaystyle\prod_{j\in a} \prod_{b\ni j, b\neq a} m_{b\to j}(x_j),\ \forall a \in \mathbb{F} \\ b_i(x_i) & = \phi_i(x_i) \exp(\dfrac{1}{d_i - 1} - \gamma_i) \displaystyle\prod_{b\ni i} m_{a\to i}(x_i),\ \forall i \in \mathbb{V} \end{cases}$$

with the (invertible) parametrization

$$\lambda_{ai}(x_i) = \log \prod_{b\ni i, b\neq a} m_{b\to i}(x_i),$$

Enforcing constraints (2.14) yields the BP fixed points equations with normalization terms γ_i. We will return to this variational setting in Section 6.

3 Normalization and existence of fixed points

We discuss here an aspect of the algorithm that did not get that much attention in the literature, which is the equivalence of the fixed points of the normalized and plain BP flavors.

It is not immediate to check that the normalized version of the algorithm does not introduce new fixed points, that would therefore not correspond to true stationary points of the Bethe free energy. We show in Theorem 3.2

that the sets of fixed points are equivalent, except possibly when the graph $\mathcal{G}$ has one unique cycle.

As pointed out by Mooij and Kappen (2007), many different sets of messages can correspond to the same set of beliefs. The following lemma shows that the set of messages leading to the same beliefs is simply constructed through linear mappings.

Lemma 3.1. *Two set of messages m and m' lead to the same beliefs if, and only if, there is a set of strictly positive constants c_{ai} such that*

$$m'_{a\to i}(x_i) = c_{ai} m_{a\to i}(x_i).$$

Proof. The direct part of the lemma is trivial. Concerning the other part, we have from (2.3) and (2.4)

$$\frac{b_a(\mathbf{x}_a) Z_a(m)}{\psi_a(\mathbf{x}_a)} = \prod_{j\in a} \prod_{b\ni j, b\neq a} m_{b\to j}(x_j)$$

$$\frac{b_i(x_i) Z_i(m)}{\phi_i(x_i)} = \prod_{a\ni i} m_{a\to i}(x_i).$$

Assume the two vectors of messages m and m' lead to the same set of beliefs b and write $m_{a\to i}(x_i) = c_{ai,x_i}\, m'_{a\to i}(x_i)$. Then, from the relation on b_i, the vector $\mathbf{c}$ satisfies

$$\prod_{a\ni i} c_{ai,x_i} = \prod_{a\ni i} \frac{m_{a\to i}(x_i)}{m'_{a\to i}(x_i)} = \frac{Z_i(m)}{Z_i(m')} \overset{\text{def}}{=} v_i. \tag{3.1}$$

Moreover, we want to preserve the beliefs b_a. Using (3.1), we have

$$\prod_{j\in a} \frac{m_{a\to j}(x_j)}{m'_{a\to j}(x_j)} = \prod_{j\in a} c_{aj,x_j} = \frac{Z_a(m')}{Z_a(m)} \prod_{i\in a} v_i \overset{\text{def}}{=} v_a, \tag{3.2}$$

Since v_i (resp. v_a) does not depend on the choice of x_i (resp. $\mathbf{x}_a$), (3.2) implies the independence of c_{ai,x_i} with respect to x_i. Indeed, if we compare two vectors $\mathbf{x}_a$ and $\mathbf{x}'_a$ such that, for all $i \in a \setminus j$, $x'_i = x_i$, but $x'_j \neq x_j$, then $c_{aj,x_j} = c_{aj,x'_j}$, which concludes the proof. ■

3.1 From normalized BP to plain BP

We show that in most cases the fixed points of a normalized BP algorithm (no matter the normalization used) are associated with fixed points of the plain BP algorithm. Recall that C is the number of independent cycles of $\mathcal{G}$.

Theorem 3.2. *A fixed point $\tilde{m}$ of the BP algorithm with normalized messages corresponds to a fixed point of the plain BP algorithm associated to the same beliefs iff one of the two following conditions is satisfied:*

(i) the graph $\mathcal{G}$ has either no cycle or more than one ($C \neq 1$);

(ii) $C = 1$, and the normalization constants of the associated beliefs are such that

$$\prod_{a \in \mathbb{F}} Z_a(\tilde{m}) \prod_{i \in \mathbb{V}} Z_i(\tilde{m})^{1-d_i} = 1. \tag{3.3}$$

Proof. Let $\tilde{m}$ be a fixed point of (2.8). Let us find a set of constants c_{ai} such that $m_{a\to i}(x_i) = c_{ai}\, \tilde{m}_{a\to i}(x_i)$ is a non-zero fixed point of (2.1, 2.2). Using Lemma 3.1, we see that m and $\tilde{m}$ correspond to the same beliefs. We have

$$\begin{aligned}
\Theta_{ai,x_i}(m) &= \left[\prod_{j\in a\setminus i} \prod_{a'\ni j, a'\neq a} c_{a'j}\right] \Theta_{ai,x_i}(\tilde{m}) \\
&= \left[\prod_{j\in a\setminus i} \prod_{a'\ni j, a'\neq a} c_{a'j}\right] Z_{ai}\, \tilde{m}_{a\to i}(x_i) \\
&= \frac{1}{c_{ai}}\left[\prod_{j\in a\setminus i} \prod_{a'\ni j, a'\neq a} c_{a'j}\right] Z_{ai}\, m_{a\to i}(x_i),
\end{aligned}$$

and therefore

$$\log c_{ai} - \sum_{j\in a\setminus i} \sum_{a'\ni j, a'\neq a} \log c_{a'j} = \log Z_{ai}.$$

This equation is precisely in the setting of Lemma A.2 given in the Appendix, with $x_{ai} = \log c_{ai}$ and $y_{ai} = \log Z_{ai} = \log Z_a - \log Z_i$. It always has a solution when $C \neq 1$; when $C = 1$, the additional condition (A.5) is required, and (3.3) follows. ■

There is in general an infinite number of fixed points m corresponding to each $\tilde{m}$. However, as noted at the beginning of the section, this is not a problem, since all these fixed points correspond to the same set of beliefs. In this sense, normalizing the messages can have the effect of collapsing equivalent fixed points.

When $C = 1$, it is known (Weiss, 2000) that normalized BP always converges to a fixed point. However, the theorem above states that there may be no basic fixed point m corresponding to a given $\tilde{m}$.

It is actually not difficult to see what happens in this case: assume a trivial network with two variables and two factors $a = b = \{1, 2\}$ and assume for simplicity that $\phi_1 = \phi_2 = 1$. The equations for the BP fixed point boil down to relations like

$$m_{a\to 1}(x_i) = \sum_{x_2} \psi_a(x_1, x_2) m_{b\to 2}(x_2),$$

or, with a matrix notation,

$$\mathbf{m}_{a\to 1} = \Psi_a \mathbf{m}_{b\to 2} = \Psi_a \Psi_b \mathbf{m}_{a\to 1}.$$

Therefore, the matrix $\Psi_a\Psi_b$ necessarily has 1 as an eigenvalue. Since this is not true in general, there can be no fixed point for basic BP. In the normalized case, Weiss (2000) shows that BP always converges to the Perron vector of this matrix. We know there is an infinite number (not even countable, see Lemma 3.1) set of messages corresponding to the same beliefs.

It is possible that the behavior of the algorithm leads to convergence of the beliefs without the convergence of messages as the case $C = 1$ suggests. Indeed, the plain BP scheme is then a linear dynamical system which can converge to a subspace as described in Hartfiel (1997). We will describe more precisely this kind of behavior in Section 4.

3.2 From plain BP to normalized BP

It turns out that there is no general result about whether a plain BP fixed point is mapped to a fixed point by normalization. In this section, we will thus first examine the case of a fairly general family of normalizations, and then look at two other examples.

Definition 3.3. A normalization Z_{ai} is said to be *positive homogeneous* when it is of the form $Z_{ai} = N_{ai} \circ \Theta_{ai}$, with $N_{ai} : \mathbb{R}^q \mapsto \mathbb{R}$ positive homogeneous functions of order 1 satisfying

$$N_{ai}(\lambda m_{a\to i}) = \lambda N_{ai}(m_{a\to i}), \forall \lambda \geq 0. \tag{3.4}$$

$$N_{ai}(m_{a\to i}) = 0 \iff m_{a\to i} = 0. \tag{3.5}$$

The part $\Longleftarrow$ of (3.5) is obviously implied by (3.4). A particular family of positive homogeneous normalizations is built from all norms N_{ai} on $\mathbb{R}^q$. These contain in particular the normalization $Z_{ai}^{\text{mess}}(m)$ (2.9) or the maximum of messages

$$Z_{ai}^{\infty}(m) \stackrel{\text{def}}{=} \max_x \Theta_{ai,x}(m).$$

It is actually not necessary to have a proper norm: Watanabe and Fukumizu (2009) use a scheme that amounts to

$$Z_{ai}^{1}(m) \stackrel{\text{def}}{=} \Theta_{ai,1}(m).$$

The following proposition describes the effect of the above family of normalizations.

Proposition 3.4. *All the fixed points of the plain BP algorithm leading to the same set of beliefs correspond to a unique fixed point of a positive homogeneous normalized scheme.*

Proof. Let m be a fixed point of the plain BP scheme. Using Lemma 3.1, a fixed point $\tilde{m}$ of the normalized scheme associated with the same beliefs than m is such as

$$\tilde{m}_{a\to i}(x_i) = c_{ai} m_{a\to i}(x_i). \tag{3.6}$$

Since Θ is multi-linear,

$$\Theta_{ai,x_i}(\tilde{m}) = \left(\prod_{j\in a\setminus i} \prod_{d\ni j, d\neq a} c_{dj} \right) \Theta_{ai,x_i}(m),$$

and, using (3.4),

$$Z_{ai}(\tilde{m}) = \left(\prod_{j\in a\setminus i} \prod_{d\ni j, d\neq a} c_{dj} \right) Z_{ai}(m),$$

$$\tilde{m}_{a\to i}(x_i) = \frac{\Theta_{ai,x_i}(\tilde{m})}{Z_{ai}(\tilde{m})} = \frac{m_{a\to i}(x_i)}{Z_{ai}(m)}.$$

Therefore, $\tilde{m}$ is determined uniquely from m. Since $\tilde{m}$ is clearly invariant for all the set of messages m corresponding to the same beliefs (see Lemma 3.1), the proof is complete. ■

In order to emphasize the result of Proposition 3.4, it is interesting to describe what happens with the belief normalization Z^{bel} (2.11). We know from Lemma 2.1 that, for any normalization, we have at any fixed point

$$Z_{ai}(m) = \frac{Z_a(m)}{Z_i(m)} \overset{\text{def}}{=} Z_{ai}^{\text{bel}}(m).$$

Therefore, any fixed point of any normalized scheme (even of the plain scheme) is a fixed point of the scheme with normalization Z^{bel}. We see the difference between this kind of normalization and a positive homogeneous one. While the latter collapses families of fixed points to one unique fixed point, Z^{bel} instead conserves all the fixed points of all possible schemes.

To conclude this section, we will present an example of a "bad normalization" to illustrate a worst case scenario. Consider the following normalization

$$Z_{ai}(m) = \frac{\sum_x \Theta_{ai,x}(m)}{\sup_x m_{a\to i}(x)}.$$

This normalization, which is not homogeneous at all, defines a BP algorithm which does not admit any fixed point. Following the proof of Proposition 3.4, let $\tilde{m}$ be a fixed point of normalized BP associated with a plain fixed point m through (3.6), then

$$\tilde{m}_{a\to i}(x_i) = \frac{\Theta_{ai,x_i}(\tilde{m})}{Z_{ai}(\tilde{m})} = \frac{\tilde{m}_{a\to i}(x_i)}{Z_{ai}(m)}$$

Indeed it is easy to check that

$$Z_{ai}(\tilde{m}) = \frac{\prod_{j\in a\setminus i}\prod_{b\ni j,b\neq a} c_{bj}}{c_{ai}} Z_{ai}(m).$$

Since for any fixed point m of the plain update we have $Z_{ai}(m) > 1$, no message $\tilde{m}$ can be a fixed point for this normalized scheme. Using Theorem 3.2 we conclude that this scheme admits no fixed point.

4 Belief dynamic

We are interested here in looking at the dynamic in terms of convergence of beliefs. At each step of the algorithm, using (2.3) and (2.4), we can compute the current beliefs $b_i^{(n)}$ and $b_a^{(n)}$ associated with the message $m^{(n)}$. The sequence $m^{(n)}$ will be said to be "b-convergent" when the sequences $b_i^{(n)}$ and $b_a^{(n)}$ converge. The term "simple convergence" will be used to refer to convergence of the sequence $m^{(n)}$ itself. Simple convergence obviously implies b-convergence. We will first show that for a positive homogeneous normalization, b-convergence and simple convergence are equivalent. We will then conclude by looking at b-convergence in a quotient space introduced in Mooij and Kappen (2007) and we show the links between these two approaches.

Proposition 4.1. *For any positive homogeneous normalization Z_{ai} with* continuous N_{ai}*, simple convergence and b-convergence are equivalent.*

Proof. Assume that the sequences of beliefs, indexed by iteration n, are such that $b_a^{(n)} \to b_a$ and $b_i^{(n)} \to b_i$ as $n \to \infty$. The idea of the proof is first to express the normalized messages $\tilde{m}_{a\to i}^{(n)}$ at each step in terms of these beliefs, and then to conclude by a continuity argument. Starting from a rewrite of (2.3)–(2.4),

$$b_i^{(n)}(x_i) = \frac{\phi_i(x_i)}{Z_i(\tilde{m}^{(n)})}\prod_{a\ni i}\tilde{m}_{a\to i}^{(n)}(x_i),$$

$$b_a^{(n)}(\mathbf{x}_a) = \frac{\psi_a(\mathbf{x}_a)}{Z_a(\tilde{m}^{(n)})}\prod_{j\in a}\phi_j(x_j)\prod_{b\ni j,b\neq a}\tilde{m}_{b\to j}^{(n)}(x_j),$$

one obtains by recombination

$$\prod_{j\in a}\tilde{m}_{a\to j}^{(n)}(x_j) = \frac{\prod_{j\in a} Z_j(\tilde{m}^{(n)})}{Z_a(\tilde{m}^{(n)})}\psi_a(\mathbf{x}_a)\frac{\prod_{j\in a} b_j^{(n)}(x_j)}{b_a^{(n)}(\mathbf{x}_a)} \overset{\text{def}}{=} \frac{K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; x_i)}{\tilde{Z}_{ai}(\tilde{m})},$$

where an arbitrary variable $i \in a$ has been singled out and

$$\frac{1}{\tilde{Z}_{ai}(\tilde{m})} \overset{\text{def}}{=} \frac{\prod_{j\in a} Z_j(\tilde{m}^{(n)})}{Z_a(\tilde{m}^{(n)})}.$$

Assume now that $\mathbf{x}_{a\backslash i}$ is fixed and consider $\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\backslash i}) \overset{\text{def}}{=} K_{ai}^{(n)}(\mathbf{x}_{a\backslash i};\cdot)$ as a vector of $\mathbb{R}^q$. Normalizing each side of the equation with a positive homogeneous function N_{ai} yields

$$\frac{\tilde{m}_{a\to i}^{(n)}(x_i)}{N_{ai}\left[\tilde{m}_{a\to i}^{(n)}\right]} = \frac{K_{ai}^{(n)}(\mathbf{x}_{a\backslash i};x_i)}{N_{ai}\left[\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\backslash i})\right]}.$$

Actually $N_{ai}\left[\tilde{m}_{a\to i}^{(n)}\right] = 1$, since $\tilde{m}_{a\to i}^{(n)}$ has been normalized by N_{ai} and therefore

$$\tilde{m}_{a\to i}^{(n)}(x_i) = \frac{K_{ai}^{(n)}(\mathbf{x}_{a\backslash i};x_i)}{N_{ai}\left[\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\backslash i})\right]}.$$

This conclude the proof, since $\tilde{m}_{a\to i}^{(n)}$ has been expressed as a continuous function of $b_i^{(n)}$ and $b_a^{(n)}$, and therefore it converges whenever the beliefs converge. ■

We follow now an idea developed in Mooij and Kappen (2007) and study the behavior of the BP algorithm in a quotient space corresponding to the invariance of beliefs. First we will introduce a natural parametrization for which the quotient space is just a vector space. Then it will be trivial to show that, in terms of b-convergence, the effect of normalization is null.

The idea of b-convergence is easier to express with the new parametrization :

$$\mu_{ai}(x_i) \overset{\text{def}}{=} \log m_{a\to i}(x_i),$$

so that the plain update mapping (2.7) becomes

$$\Lambda_{ai,x_i}(\mu) = \log\left[\sum_{\mathbf{x}_a\backslash i}\psi_a(\mathbf{x}_a)\exp\Big(\sum_{j\in a\backslash i}\sum_{\substack{b\ni j\\ b\neq a}}\mu_{bj}(x_j)\Big)\right].$$

We have $\mu \in \mathcal{N} \overset{\text{def}}{=} \mathbb{R}^{|\mathbb{E}|q}$ and we define the vector space $\mathcal{W}$ which is the linear span of the following vectors $\{e_{ai}\in\mathcal{N}\}_{(ai)\in\mathbb{E}}$

$$(e_{ai})_{cj,x_j} \overset{\text{def}}{=} \mathbb{1}_{\{ai=cj\}}.$$

It is trivial to see that the invariance set of the beliefs corresponding to μ described in Lemma 3.1 is simply the affine space $\mu+\mathcal{W}$. So the b-convergence of a sequence $\mu^{(n)}$ is simply the convergence of $\mu^{(n)}$ in the quotient space $\mathcal{N}\setminus\mathcal{W}$ (which is a vector space, see Halmos (1974)). Finally we define the notation $[x]$ for the canonical projection of x on $\mathcal{N}\setminus\mathcal{W}$.

Suppose that we resolve to some kind of normalization on μ, it is easy to see that this normalization plays no role in the quotient space. The normalization on μ leads to $\mu + w$ with some $w \in \mathcal{W}$. We have

$$\begin{aligned}\Lambda_{ai,x_i}(\mu + w) &= \log\Bigl(\sum_{j\in a\setminus i}\sum_{\substack{b\ni j\\ b\neq a}} w_{bj}\Bigr) + \Lambda_{ai,x_i}(\mu)\\ &\overset{\text{def}}{=} l_{ai} + \Lambda_{ai,x_i}(\mu),\end{aligned}$$

which can be summed up by

$$[\Lambda(\mu + \mathcal{W})] = [\Lambda(\mu)], \qquad (4.1)$$

since $l \in \mathcal{W}$. We conclude by a proposition which is directly implied by (4.1).

Proposition 4.2. *The dynamic, i.e. the value of the normalized beliefs at each step, of the BP algorithm with or without normalization is exactly the same.*

We will come back to this vision in term of quotient space in section 5.3.

5 Local stability of BP fixed points

The question of convergence of BP has been addressed in a series of works (Tatikonda and Jordan, 2002; Mooij and Kappen, 2007; Ihler et al., 2005) which establish conditions and bounds on the MRF coefficients for having global convergence. In this section, we change the viewpoint and, instead of looking for conditions ensuring a single fixed point, we examine the different fixed points for a given joint probability and their local properties.

In what follows, we are interested in the local stability of a message fixed point m with associated beliefs b. It is known that a BP fixed point is locally attractive if the Jacobian of the relevant mapping (Θ or its normalized version) at this point has a spectral radius strictly smaller than 1 and unstable when the spectral radius is strictly greater than 1. The term “spectral radius” should be understood here as the modulus of the largest eigenvalue of the Jacobian matrix.

We will first show that BP with plain messages can in fact never converge when there is more than one cycle (Theorem 5.1), and then explain how normalization of messages improves the situation (Proposition 5.2, Theorem 5.3).

5.1 Unnormalized messages

The characterization of the local stability relies on two ingredients. The first one is the oriented line graph $L(\mathcal{G})$ based on $\mathcal{G}$, which vertices are the elements of $\mathbb{E}$, and which oriented links relate ai to $a'j$ if $j \in a \cap a'$, $j \neq i$

and $a' \neq a$. The corresponding 0-1 adjacency matrix A is defined by the coefficients

$$A_{ai}^{a'j} \overset{\text{def}}{=} \mathbb{1}_{\{j \in a \cap a',\, j \neq i,\, a' \neq a\}}. \tag{5.1}$$

The second ingredient is the set of stochastic matrices $B^{(iaj)}$, attached to pairs of variables (i,j) having a factor node a in common, and which coefficients are the conditional beliefs,

$$b_{k\ell}^{(iaj)} \overset{\text{def}}{=} b_a(x_j = \ell | x_i = k) = \sum_{\mathbf{x}_{a \setminus \{i,j\}}} \frac{b_a(\mathbf{x}_a)}{b_i(x_i)}\bigg|_{\substack{x_i = k\\ x_j = \ell}}$$

for all $(k,\ell) \in \{1,\dots,q\}^2$.

Using the representation (2.10) of the BP algorithm, the Jacobian reads at this point:

$$\begin{aligned}\frac{\partial \Theta_{ai,x_i}(m)}{\partial m_{a'\to j}(x_j)} &= \sum_{\mathbf{x}_{a\setminus\{i,j\}}} \frac{b_a(\mathbf{x}_a)}{b_i(x_i)} \frac{m_{a\to i}(x_i)}{m_{a'\to j}(x_j)} \mathbb{1}_{\{j \in a\setminus i\}} \mathbb{1}_{\{a' \ni j, a' \neq a\}} \\ &= \frac{b_{ij|a}(x_i,x_j)}{b_i(x_i)} \frac{m_{a\to i}(x_i)}{m_{a'\to j}(x_j)} A_{ai}^{a'j}\end{aligned}$$

Therefore, the Jacobian of the plain BP algorithm is—using a trivial change of variable—similar to the matrix J defined, for any pair (ai,k) and $(a'j,\ell)$ of $\mathbb{E} \times \{1,\dots,q\}$ by the elements

$$J_{ai,k}^{a'j,\ell} \overset{\text{def}}{=} b_{k\ell}^{(iaj)} A_{ai}^{a'j},$$

This expression is analogous to the Jacobian encountered in Mooij and Kappen (2007). It is interesting to note that it only depends on the structure of the graph and on the belief corresponding to the fixed point.

Since $\mathcal{G}$ is a singly connected graph, it is clear that A is an irreducible matrix. To simplify the discussion, we assume in the following that J is also irreducible. This will be true as long as the ψ are always positive. It is easy to see that to any right eigenvector of A corresponds a right eigenvector of J associated to the same eigenvalue: if $\mathbf{v} = (v_{ai}, ai \in \mathbb{E})$ is such that $A\mathbf{v} = \lambda\mathbf{v}$, then the vector $\mathbf{v}^+$, defined by coordinates $v_{a'j\ell}^+ \overset{\text{def}}{=} v_{a'j}$, for all $a'j \in \mathbb{E}$ and $\ell \in \{1,\dots,q\}$, satisfies $J\mathbf{v} = \lambda\mathbf{v}$. We will say that $\mathbf{v}^+$ is a A-based right eigenvector of J. Similarly, if $\mathbf{u}$ is a left eigenvector of A, with obvious notations one can define a A-based left eigenvector $\mathbf{u}^+$ of J by the following coordinates: $u_{aik}^+ \overset{\text{def}}{=} u_{ai} b_i(k)$.

Using this correspondence between the two matrices, we can prove the following result.

Theorem 5.1. *If the graph $\mathcal{G}$ has more than one cycle ($C > 1$), and the matrix J is irreducible, then the plain BP update rules (2.1, 2.2) do not admit any stable fixed point.*

Proof. Let $\boldsymbol{\pi}$ be the right Perron vector of A, which has positive entries, since A is irreducible (Seneta, 2006, Theorem 1.5). The A-based vector $\boldsymbol{\pi}^+$ also has positive coordinates and is therefore the right Perron vector of J (Seneta, 2006, Theorem 1.6); the spectral radius of J is thus equal to the one of A.

When $C > 1$, Lemma A.1 implies that 1 is an eigenvalue of A associated to divergenceless vectors. However, such vectors cannot be non-negative, and therefore the Perron eigenvalue of A is strictly greater than 1. This concludes the proof of the theorem. ■

5.2 Positively homogeneous normalization

We have seen in Proposition 4.1 that all the continuous positively homogeneous normalizations make simple convergence equivalent to b-convergence. As a result, one expects that local stability of fixed points will again depend on the beliefs structure only. Since all the positively homogeneous normalization share the same properties, we look at the particular case of $Z_{ai}^{\mathrm{mess}}(m)$, which is both simple and differentiable. We then obtain a Jacobian matrix with more interesting properties. In particular, this matrix depends not only on the beliefs at the fixed point, but also on the messages themselves: for the normalized BP algorithm (2.8 with Z_{ai}^{mess}), the coefficients of the Jacobian at fixed point m with beliefs b read

$$\frac{\partial}{\partial \tilde{m}_{a'\to j}(\ell)}\left[\frac{\Theta_{ai,k}(\tilde{m})}{\sum_{x=1}^{q}\Theta_{ai,x}(\tilde{m})}\right]$$
$$= J_{ai,k}^{a'j,\ell}\frac{m_{a\to i}(k)}{m_{a'\to j}(\ell)} - m_{a\to i}(k)\sum_{x=1}^{q} J_{ai,x}^{a'j,\ell}\frac{m_{a\to i}(x)}{m_{a'\to j}(\ell)},$$

which is again similar to the matrix $\widetilde{J}$ of general term

$$\widetilde{J}_{ai,k}^{a'j,\ell} \overset{\mathrm{def}}{=} \left[b_{k\ell}^{(iaj)} - \sum_{x=1}^{q} m_{a\to i}(x) b_{x\ell}^{(iaj)}\right] A_{ai}^{a'j} = J_{ai,k}^{a'j,\ell} - \sum_{x=1}^{q} m_{a\to i}(x) J_{ai,x}^{a'j,\ell}. \tag{5.2}$$

It is actually possible to prove that the spectrum of $\tilde{J}$ does not depend on the messages themselves but only of the belief at the fixed point.

Proposition 5.2. *The eigenvectors of J are associated to eigenvectors of $\widetilde{J}$ with the same eigenvalues, except the A-based eigenvectors of J (including its Perron vector), which belong to the kernel of $\tilde{J}$.*

Proof. The new Jacobian matrix can be expressed from the old one as $\widetilde{J} = (\mathbb{I} - M)J$, where M is the matrix whose coefficient at row (ai, k) and column $(a'j, \ell)$ is $\mathbb{1}_{\{a=a',i=j\}} m_{a'\to j}(\ell)$. Elementary computations yield the following properties of M:

- $M^2 = M$: M is a projector;
- $\widetilde{J}M = 0$.

For any right eigenvector $\mathbf{v}$ of J associated to some eigenvalue λ,

$$\widetilde{J}(\mathbf{v} - M\mathbf{v}) = \widetilde{J}\mathbf{v} = (\mathbb{I} - M)J\mathbf{v} = \lambda(\mathbf{v} - M\mathbf{v})$$

so that $\mathbf{v} - M\mathbf{v}$ is a (right) eigenvector of $\widetilde{J}$ associated to λ, unless $\mathbf{v}$ is an A-based eigenvector, in which case $\mathbf{v} = M\mathbf{v}$ and $\mathbf{v}$ is in the kernel of $\widetilde{J}$.

Similarly, if $\mathbf{u}$ is such that $\mathbf{u}^T\widetilde{J} = \lambda\mathbf{u}^T$ for $\lambda \neq 0$, then $\lambda\mathbf{u}^T M = \mathbf{u}^T\widetilde{J}M = 0$ and therefore $\mathbf{u}^T\widetilde{J} = \mathbf{u}^T(\mathbb{I} - M)J = \mathbf{u}^T J = \lambda\mathbf{u}^T$: any non-zero eigenvalue of $\widetilde{J}$ is an eigenvalue of J. This proves the last part of the theorem. ■

As a consequence of this proposition, when J is an irreducible matrix, $\widetilde{J}$ has a strictly smaller spectral radius: the net effect of normalization is to improve convergence (although it may actually not be enough to guarantee convergence). To quantify this improvement of convergence related to message normalization, we resort to classical arguments used in speed convergence of Markov chains (see e.g. Brémaud (1999)).

The presence of the messages in the Jacobian matrix $\widetilde{J}$ complicates the evaluation of this effect. However, it is known (see e.g. Furtlehner et al. (2010)) that it is possible to chose the functions $\hat{\phi}$ and $\hat{\psi}$ as

$$\hat{\phi}_i(x_i) \stackrel{\text{def}}{=} \hat{b}_i(x_i), \qquad \hat{\psi}_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \frac{\hat{b}_a(\mathbf{x}_a)}{\prod_{i\in a}\hat{b}_i(x_i)}, \tag{5.3}$$

in order to obtain a prescribed set of beliefs $\hat{b}$ at a fixed point. Indeed, BP will admit a fixed point with $b_a = \hat{b}_a$ and $b_i = \hat{b}_i$ when $m_{a\to i}(x_i) \equiv 1$. Since only the beliefs matter here, without loss of generality, we restrict ourselves in the remainder of this section to the functions (5.3). Then, from (5.2), the definition of $\widetilde{J}$ rewrites

$$\widetilde{J}_{ai,k}^{a'j,\ell} \stackrel{\text{def}}{=} \left[b_{k\ell}^{(iaj)} - \frac{1}{q}\sum_{x=1}^{q} b_{x\ell}^{(iaj)}\right] A_{ai}^{a'j} = J_{ai,k}^{a'j,\ell} - \frac{1}{q}\sum_{x=1}^{q} J_{ai,x}^{a'j,\ell}.$$

For each connected pair (i, j) of variable nodes, we associate to the stochastic kernel $B^{(iaj)}$ a combined stochastic kernel $K^{(iaj)} \stackrel{\text{def}}{=} B^{(iaj)}B^{(jai)}$, with coefficients

$$K_{k\ell}^{(iaj)} \stackrel{\text{def}}{=} \sum_{m=1}^{q} b_{km}^{(iaj)} b_{m\ell}^{(jai)}. \tag{5.4}$$

Since $b^{(i)}B^{(iaj)} = b^{(j)}$, $b^{(i)}$ is the invariant measure associated to K:

$$b^{(i)}K^{(iaj)} = b^{(i)}B^{(iaj)}B^{(jai)} = b^{(j)}B^{(jai)} = b^{(i)}$$

and $K^{(iaj)}$ is reversible, since

$$\begin{aligned} b_k^{(i)} K_{k\ell}^{(iaj)} &= \sum_{m=1}^{q} b_{mk}^{(jai)} b_m^{(j)} b_{m\ell}^{(jai)} \\ &= \sum_{m=1}^{q} b_{mk}^{(jai)} b_{\ell m}^{(iaj)} b_\ell^{(i)} = b_\ell^{(i)} K_{\ell k}^{(iaj)}. \end{aligned}$$

Let $\mu_2^{(iaj)}$ be the second largest eigenvalue of $K^{(iaj)}$ and let

$$\mu_2 \overset{\text{def}}{=} \max_{ij} |\mu_2^{(iaj)}|^{\frac{1}{2}}.$$

The combined effect of the graph and of the local correlations, on the stability of the reference fixed point is stated as follows.

Theorem 5.3. *Let λ_1 be the Perron eigenvalue of the matrix A*

(i) if $\lambda_1 \mu_2 < 1$, the fixed point of the normalized BP schema (2.8 with Z_{ai}^{mess}) associated to b is stable.

(ii) condition (i) is necessary and sufficient if the system is homogeneous ($B^{(iaj)} = B$ independent of i, j and a), with μ_2 representing the second largest eigenvalue of B.

Proof. See Appendix B ■

The quantity μ_2 is representative of the level of mutual information between variables. It relates to the spectral gap (see e.g. Diaconis and Strook (1991) for geometric bounds) of each elementary stochastic matrix $B^{(iaj)}$, while λ_1 encodes the statistical properties of the graph connectivity. The bound $\lambda_1 \mu_2 < 1$ could be refined when dealing with the statistical average of the sum over path in (B.1) which allows to define μ_2 as

$$\mu_2 = \lim_{n\to\infty} \max_{(ai,a'j)} \Big\{ \frac{1}{|\Gamma_{ai,a'j}^{(n)}|} \sum_{\gamma \in \Gamma_{ai,a'j}^{(n)}} \Big(\prod_{(x,y)\in\gamma} \mu_2^{(xy)} \Big)^{\frac{1}{2n}} \Big\}.$$

5.3 Local convergence in quotient space $\mathcal{N} \setminus \mathcal{W}$

The idea is to make the connexion between local stability of fixed point as described previously and the same notion of local stability but in the quotient space $\mathcal{N} \setminus \mathcal{W}$ described in Section 4. Trivial computation based on the results of Section 5.1 gives us the derivatives of Λ.

$$\frac{\partial \Lambda_{ai,x_i}(\mu)}{\partial \mu_{bj}(x_j)} = \frac{b_{ij|a}(x_i,x_j)}{b_i(x_i)} A_{ai}^{bj} = J_{ai,x_i}^{bj,x_j}.$$

In terms of convergence in $\mathcal{N}\backslash\mathcal{W}$, the stability of a fixed point is given by the projection of J on the quotient space $\mathcal{N}\backslash\mathcal{W}$ and we have (Mooij and Kappen, 2007) :

$$[J] \overset{\text{def}}{=} [\nabla\Lambda] = \nabla[\Lambda]$$

Proposition 5.4. *The eigenvalues of $[J]$ are the eigenvalues of J which are not associated with A-based eigenvectors. The A-based eigenvectors of J belong to the kernel of $[J]$*

Proof. Let v be an eigenvector of J for the eigenvalue λ, we have

$$[Jv] = [\lambda v] = \lambda[v],$$

so $[v]$ is an eigenvector of $[J]$ with the same eigenvalue λ iff $[v] \neq 0$. The A-based eigenvectors (see Section 5.1) w of J belongs to $\mathcal{W}$ so we have

$$[w] = 0.$$

It means that these eigenvectors of J have no equivalent w.r.t $[J]$ and play no role in belief fixed point stability. ■

We have seen that the normalization Z_{ai}^{mess} is equivalent to multiplying the jacobian matrix J by the projection $\mathbb{I} - M$ (Proposition 5.2), with

$$\ker(\mathbb{I} - M) = \mathcal{W}.$$

The projection $\mathbb{I} - M$ is in fact a quotient map from $\mathcal{N}$ to $\mathcal{N} \setminus \mathcal{W}$. So the normalization Z_{ai}^{mess} is strictly equivalent, when we look at the messages $m_{a\to i}(x_i)$, to working on the quotient space $\mathcal{N} \setminus \mathcal{W}$. More generally for any differentiable positively homogeneous normalization we will obtain the same result, the jacobian of the corresponding normalized scheme will be the projection of the jacobian J on the quotient space $\mathcal{N} \setminus \mathcal{W}$, through some quotient map.

6 Normalization in the variational problem

Since Proposition 4.2 shows that the choice of normalization has no real effect on the dynamic of BP, it will have no effect on b-convergence either. In this section, we turn to the effect of normalization on the underlying variational problem. It will be assumed here that the beliefs b_i and b_a are normalized (2.5) and compatible (2.14). If only (2.14) is satisfied, they will be denoted β_i and β_a. It is quite obvious that imposing only compatibility constraints leads to a unique normalization constant Z

$$Z(\beta) \overset{\text{def}}{=} \sum_{x_i} \beta_i(x_i) = \sum_{\mathbf{x}_a} \beta_a(\mathbf{x}_a),$$

which is not *a priori* related to the constants $Z_a(m)$ and $Z_i(m)$ seen in the previous sections. The quantities $\beta_i(x_i)/Z(\beta)$ and $\beta_a(\mathbf{x}_a)/Z(\beta)$ can be denoted as $b_i(x_i)$ and $b_a(\mathbf{x}_a)$ since (2.5) holds for them.

The aim of this section is to explicit the relationship between the minimizations of the Bethe free energy (2.15) with and without normalization constraints (2.5). Generally speaking, we can express them as a minimization problem $\mathcal{P}(E)$ on some set E as

$$\mathcal{P}(E) \quad : \quad \operatorname*{argmin}_{\beta\in E} F(\beta) \tag{6.1}$$

where E is chosen as follows

- plain case: $E = E_1$ is the set of positive measures such as (2.14) holds,
- normalized case: $E = E_2 \subsetneq E_1$ has the additional constraint (2.5).

It is possible to derive a BP algorithm for the plain problem following the same path as in Section 2. The resulting update equations will be identical, except for the γ_i terms.

The first step is to compare the solutions of (6.1) on E_1 and E_2. Let φ be the bijection between E_1 and $E_2 \times \mathbb{R}_+^*$,

$$\begin{aligned} \varphi : E_2 \times \mathbb{R}_+^* &\longrightarrow E_1 \\ (b, Z) &\longrightarrow bZ. \end{aligned}$$

The variational problem $\mathcal{P}(E_1)$ is equivalent to

$$(\hat{b}, \hat{Z}) = \operatorname*{argmin}_{(b,Z)\in E_2} F(\varphi(b, Z)),$$

with $\varphi(\hat{b}, \hat{Z}) = \hat{b}\hat{Z} = \hat{\beta} \overset{\text{def}}{=} \operatorname*{argmin}_{\beta\in E_1} F(\beta)$.

The next step is to express the Bethe free energy $F(\beta)$ of an unnormalized positive measure β as a function of the Bethe free energy of the corresponding normalized measure b.

Lemma 6.1. *As soon as the factor graph is connected, for any $\beta = Zb \in E_1$ we have*

$$F(Zb) = Z\big(F(b) + (1 - C)\log Z\big), \tag{6.2}$$

with C being the number of independent cycles of the graph.

Proof.

$$\begin{aligned} F(\beta) &= F(Zb) \\ &= Z\Big[\sum_{a,\mathbf{x}_a} b_a(\mathbf{x}_a)\log\Big(\frac{Zb_a(\mathbf{x}_a)}{\psi_a(\mathbf{x}_a)}\Big) + \sum_{i,x_i} b_i(x_i)\log\Big(\frac{(Zb_i(x_i))^{1-d_i}}{\phi_i(x_i)}\Big)\Big] \\ &= Z\Big(F(b) + (|\mathbb{F}| + |\mathbb{V}| - |\mathbb{E}|)\log Z\Big) \\ &= Z\Big(F(b) + (1 - C)\log Z\Big), \end{aligned}$$

where the last equality comes from elementary graph theory (see e.g. Berge (1967)). ■

The quantity $1 - C$ will be negative in the nontrivial cases (at least 2 cycles). Since all the Zb are equivalent from our point of view, we look at the derivatives of $F(Zb)$ as a function of Z to see what happens in the plain variational problem.

Theorem 6.2. *The normalized beliefs corresponding to the extrema of the plain variational problem $\mathcal{P}(E_1)$ are exactly the same as the ones of the normalized problem $\mathcal{P}(E_2)$ as soon as $C \neq 1$.*

Proof. Using Lemma 6.1 we obtain

$$\frac{\partial F(\beta)}{\partial Z} = F(b) + (1 - C)(\log Z + 1),$$

the stationary points are

$$\hat{Z} = \exp\Big(\frac{F(b)}{C-1} - 1\Big). \tag{6.3}$$

At these points we can compute the Bethe free energy

$$F(\hat{\beta}) = F(\hat{Z}b) = (C-1)\exp\Big(\frac{F(b)}{C-1} - 1\Big) = G(F(b)).$$

It is easy to check that, if $C \neq 1$, G is an increasing function, so the extrema of $F(\beta)$ are reached at the same normalized beliefs. More precisely, if b_1 and b_2 are elements of E_2 such that $F(b_1) \leq F(b_2)$ then $F(\hat{\beta}_1 = \hat{Z}_1 b_1) \leq F(\hat{\beta}_2 = \hat{Z}_2 b_2)$, which allows us to conclude. ■

In other words, imposing a normalization in the variational problem or normalizing after a solution is reached is equivalent as long as $C \neq 1$. Moreover, in the unnormalized case, the Bethe free energy at the local extremum writes

$$F(b) = (\mathcal{C} - 1)(\log \hat{Z} + 1). \tag{6.4}$$

We can therefore compare the "quality" of different fixed points by comparing only the normalization constant obtained: the smaller Z is, the better the approximation, modulo the fact that we're not minimizing a true distance.

When $C = 1$, it has been shown already in Section 3 that the normalized scheme is always convergent, whereas the plain scheme can have no fixed point. In this case, (6.2) rewrites

$$F(\beta) = F(Zb) = ZF(b).$$

The form of this relationship shows what happens: if the extremum of the normalized variational problem is strictly negative, $F(\beta)$ is unbounded from below and Z will diverge to $+\infty$; conversely, if the extremum is strictly positive, Z will go to zero. In the (very) particular case where the minimum of the normalized problem is equal to zero, the problem is still well defined. In fact this condition $F(b) = 0$ is equivalent to the one of Theorem 3.2 when $\mathcal{C} = 1$.

To sum up, as soon as the plain variational problem is well defined, it is equivalent to the normalized one and the normalization constant allows to compute easily the Bethe free energy using (6.4). When this is no longer the case, we still know that the dynamics of both algorithms remain the same (Proposition 4.2) but the plain variational problem (which can still converge in terms of beliefs) will not converge in terms of normalization constant Z, and we have no more easy information on the fixed point free energy.

As emphasized previously, the relationship between $\hat{Z}$, $Z_a(m)$ and $Z_i(m)$ is not trivial. In the case of the plain BP algorithm, for which $Z_a(m) = Z_i(m)$, an elementary computation yields the following relation at any fixed point

$$F(b) = (\mathcal{C} - 1) \log Z_a(m),$$

which seemingly contradicts (6.4). In fact, the algorithm derived from the plain variational problem is not exactly the plain BP scheme. Usually, since one resorts to some kind of normalization, the multiplicative constants of the fixed point equations are discarded (see Yedidia et al. (2005) for more details). Keeping track of them yields

$$m_{a\to i}(x_i) = \exp\left(\frac{d_i - 2}{d_i - 1}\right) \Theta_{ai,x_i}(m), \tag{6.5}$$

$$\beta_a(\mathbf{x}_a) = \frac{1}{e}\psi_a(\mathbf{x}_a) \prod_{j\in a} n_{j\to a}(x_j),$$

$$\beta_i(x_i) = \phi_i(x_i) \exp\left(\frac{1}{d_i - 1}\right) \prod_{b\ni i} m_{b\to i}(x_i).$$

Actually, the plain update scheme (2.1,2.2) corresponds to some constant normalization $\exp\left(\frac{d_i-2}{d_i-1}\right)$. Without any normalization, using (6.5) as update rule, one would obtain

$$\hat{Z} = \frac{Z_a(m)}{e} = Z_i(m) \exp\left(\frac{1}{d_i - 1}\right).$$

7 Conclusion

This paper motivation was to fill a void in the literature about the effect of normalization on the BP algorithm. What we have learnt can be summarized in a few main points

- using a normalization in BP can in some rare cases kill or create new fixed points;
- not all normalizations are created equal when it comes to message convergence, but there is a big category of positive homogeneous normalization that all have the same effect;
- the user is ultimately concerned with convergence of beliefs, and thankfully the dynamic of normalized beliefs is insensitive to normalization.

The messages having no interest by themselves, it is worthy of remark that combining the update rules (2.12) recalled below

$$m_{a\to i}(x_i) \leftarrow \frac{b_{i|a}(x_i)}{b_i(x_i)} m_{a\to i}(x_i),$$

and the definition (2.3) and (2.4) of beliefs, one can eliminate the messages and obtain

$$\begin{aligned} b_i(x_i) &\leftarrow b_i(x_i) \prod_{a\ni i} \frac{b_{i|a}(x_i)}{b_i(x_i)}, \\ b_a(\mathbf{x}_a) &\leftarrow b_a(\mathbf{x}_a) \prod_{i\in a} \prod_{c\ni i, c\neq a} \frac{b_{i|c}(x_i)}{b_i(x_i)}, \end{aligned}$$

One particularity of these update rules is that they do not depend on the functions ψ or ϕ but only on the graph structure. The dependency on the joint law (1.1) occurs only through the initial conditions. This "product sum" algorithm therefore shares common properties for all models build on the same underlying graph, and the initial conditions should impose the details of the joint law. To our knowledge this algorithm has never been studied and we let it for future work.

A Spectral properties of the factor graph

This appendix is devoted to some properties of the matrix A defined in (5.1) that are used in Sections 3 and 5.

We consider two types of fields associated to $\mathcal{G}$, namely scalar fields and vector fields. Scalar fields are quantities attached to the vertices of the graph, while vector fields are attached to its edges. A vector field $\mathbf{w} = \{w_{ai},\ ai \in \mathbb{E}\}$ is *divergenceless* if

$$\forall a \in \mathbb{F},\ \sum_{i\in a} w_{ai} = 0 \quad \text{and} \quad \forall i \in \mathbb{V},\ \sum_{a\ni i} w_{ai} = 0.$$

A vector field $\mathbf{u} = \{u_{ai},\ ai \in \mathbb{E}\}$ is a *gradient* if there exists a scalar field $\{u_a, u_i,\ a \in \mathbb{F},\ i \in \mathbb{V}\}$ such that

$$\forall ai \in \mathbb{E},\ u_{ai} = u_a - u_i.$$

There is an orthogonal decomposition of any vector field into a divergenceless and a gradient component. Indeed, the scalar product

$$\mathbf{w}^T\mathbf{u} = \sum_{ai\in\mathbb{E}} w_{ai}u_{ai} = \sum_{a\in\mathbb{F}} u_a \sum_{i\in a} w_{ai} - \sum_{i\in\mathbb{V}} u_i \sum_{a\ni i} w_{ai},$$

is 0 for all gradient fields $\mathbf{u}$ iff $\mathbf{w}$ is divergenceless. Dimensional considerations show that any vector field $\mathbf{v}$ can be decomposed in this way.

In the following, it will be useful to define the Laplace operator Δ associated to $\mathcal{G}$. For any scalar field $\mathbf{u}$:

$$(\Delta\mathbf{u})_a \overset{\text{def}}{=} d_a u_a - \sum_{i\in a} u_i, \qquad \forall a \in \mathbb{F} \tag{A.1}$$

$$(\Delta\mathbf{u})_i \overset{\text{def}}{=} d_i u_i - \sum_{a\ni i} u_a, \qquad \forall i \in \mathbb{V}. \tag{A.2}$$

The following lemma describes the spectrum of A in terms of a Laplace equation on the graph $\mathcal{G}$.

Lemma A.1. *(i) Both gradient and divergenceless vector spaces are A-invariant and divergenceless vectors are eigenvectors of A with eigenvalue 1. (ii) eigenvectors associated to eigenvalues $\lambda \neq 1$ are gradient vectors of a scalar field $\mathbf{u}$ which satisfies*

$$\big(\Delta\mathbf{u}\big)_a = \frac{(\lambda-1)(d_a-1)}{\lambda}u_a \ \textit{and}\ \big(\Delta\mathbf{u}\big)_i = (1-\lambda)u_i. \tag{A.3}$$

and there exists a gradient vector associated to 1 iff $\mathcal{G}$ has exactly one cycle ($C = 1$).

Proof. The action of A on a given vector $\mathbf{x}$ reads

$$\sum_{a'j\in\mathbb{E}} A_{ai}^{a'j} x_{a'j} = \sum_{j\in a}\Big(\sum_{a'\ni j} x_{a'j} - x_{aj}\Big) - \sum_{a'\ni i} x_{a'i} + x_{ai},$$

The first two terms in the second member vanish if $\mathbf{x}$ is divergenceless. In addition, the first term in parentheses is independent of i while the second one is independent of a so the first assertion is justified. We concentrate then on solving the eigenvalue equation $A\mathbf{x} - \lambda\mathbf{x} = 0$ for a gradient vector $\mathbf{x}$, with $x_{ai} = u_a - u_i$. $A\mathbf{x} - \lambda\mathbf{x}$ is the gradient of a constant scalar $K \in \mathbb{R}$, and by identification we have

$$\begin{cases} \big(\Delta\mathbf{u}\big)_a + \sum_{j\in a}\big(\Delta\mathbf{u}\big)_j = (1-\lambda)u_a + K \\ \big(\Delta\mathbf{u}\big)_i = (1-\lambda)u_i + K. \end{cases}$$

The Laplacian of a constant scalar is zero, so for $\lambda \neq 1$, K may be reabsorbed in $\mathbf{u}$ and, combining these two equations with the help of identities (A.1,A.2), yields equation (A.3). For $\lambda = 1$, we obtain

$$\big(\Delta\mathbf{u}\big)_a = (1-d_a)K \qquad \text{and} \qquad \big(\Delta\mathbf{u}\big)_i = K. \tag{A.4}$$

Let D be the diagonal matrix associated to the graph $\mathcal{G}$, whose diagonal entries are the degrees d_a and d_i of each node. $M = \mathbb{I} - D^{-1}\Delta$ is a stochastic irreducible matrix, which unique right Perron vector $(1,\dots,1)$ generates the kernel of Δ. As a result, for $K = 0$, the solution to (A.4) is $u_a = u_i = cte$ so that $x_{ai} = 0$.

For $K \neq 0$, there is a solution if the second member of (A.4) is orthogonal (Δ is a symmetric operator) to the kernel. The condition reads

$$0 = \sum_a (1-d_a) + \sum_i 1 = |\mathbb{F}| - |\mathbb{E}| + |\mathbb{V}| = 1 - C,$$

where the last equality comes from elementary graph theory (see e.g. Berge (1967)). ■

Since 1 is an eigenvalue of A, it is interesting to investigate linear equations involving $\mathbb{I} - A$. Since it is already known that divergenceless vectors are in the kernel of this matrix, we restrict ourselves to the case where the constant term is of gradient type.

Lemma A.2. *For a given gradient vector field* $\mathbf{y}$*, the equation*

$$\big(\mathbb{I} - A\big)\mathbf{x} = \mathbf{y},$$

has a solution (unique up to a divergenceless vector) iff $C \neq 1$ *or* $C = 1$ *and*

$$\sum_{a\in\mathbb{F}} y_a + \sum_{i\in\mathbb{V}} (1-d_i)y_i = 0. \tag{A.5}$$

Proof. We look here only for gradient-type solutions $x_{ai} = u_a - u_i$ and write $y_{ai} = y_a - y_i$. Owing to the same arguments as in Lemma A.1, there exists a constant K such that

$$\big(\Delta\mathbf{u}\big)_a = K(d_a - 1) + y_a - \sum_{j\in a} y_j$$
$$\big(\Delta\mathbf{u}\big)_i = y_i - K.$$

Stating as before the compatibility condition for this equation yields

$$\sum_{a\in\mathbb{F}} y_a + \sum_{i\in\mathbb{V}}(1-d_i)y_i = K(C-1).$$

It is always possible to find a suitable K as long as $C \neq 1$ and when $C = 1$, (A.5) has to hold. ■

B Proof of Theorem 5.3

Let us start with (ii): when the system is homogeneous, $\widetilde{J}$ is a tensor product of A with $\widetilde{B}$, and its spectrum is therefore the product of their respective spectra. In particular if $\mathcal{G}$ has uniform degrees d_a and d_i, the condition reads

$$\mu_2(d_a-1)(d_i-1) < 1.$$

In order to prove part (i) of the theorem, we will consider a local norm on $\mathbb{R}^q$ attached to each variable node i,

$$\|x\|_{b^{(i)}} \overset{\text{def}}{=} \Bigl(\sum_{k=1}^{q} x_k^2 b_k^{(i)}\Bigr)^{\frac{1}{2}} \text{ and } \langle x\rangle_{b^{(i)}} \overset{\text{def}}{=} \sum_{k=1}^{q} x_k b_k^{(i)},$$

the local average of $x \in \mathbb{R}^q$ w.r.t $b^{(i)}$. For convenience we will also consider the somewhat hybrid global norm on $\mathbb{R}^{q\times|\mathbb{E}|}$

$$\|x\|_{\pi,b} \overset{\text{def}}{=} \sum_{a\to i} \pi_{ai}\|x_{ai}\|_{b^{(i)}},$$

where $\boldsymbol{\pi}$ is again the right Perron vector of A, associated to λ_1.

We have the following useful inequality.

Lemma B.1. *For any $(x_i, x_j) \in \mathbb{R}^{2q}$, such that $\langle x_i\rangle_{b^{(i)}} = 0$ and $x_{j,\ell} b_\ell^{(j)} = \sum_k x_{i,k} b_k^{(i)} B_{k\ell}^{(iaj)}$,*

$$\langle x_j\rangle_{b^{(j)}} = 0 \qquad \textit{and} \qquad \|x_j\|_{b^{(j)}}^2 \leq \mu_2^{(iaj)}\|x_i\|_{b^{(i)}}^2.$$

Proof. By definition (5.4), we have

$$\begin{aligned}\|x^{(j)}\|_{b^{(j)}}^2 &= \sum_{k=1}^{q} \frac{1}{b_k^{(j)}}\Bigl|\sum_{\ell=1}^{q} b_{\ell k}^{(iaj)} b_\ell^{(i)} x_\ell^{(i)}\Bigr|^2 \\ &= \sum_{\ell,m} x_\ell^{(i)} x_m^{(i)} K_{\ell m}^{(iaj)} b_\ell^{(i)}.\end{aligned}$$

Since $K^{(iaj)}$ is reversible we have from Rayleigh's theorem

$$\mu_2^{(iaj)} \overset{\text{def}}{=} \sup_x \Bigl\{\frac{\sum_{k\ell} x_k x_\ell K_{k\ell}^{(iaj)} b_k^{(i)}}{\sum_k x_k^2 b_k^{(i)}}, \langle x\rangle_{b^{(i)}} = 0, x \neq 0\Bigr\},$$

which concludes the proof. ■

To deal with iterations of J, we express it as a sum over paths.

$$\big(J^n\big)_{ai,k}^{a'j,\ell} = \big(A^n\big)_{ai}^{a'j}\big(B^{(n)}_{ai,a'j}\big)_{k\ell},$$

where $B^{(n)}_{ai,a'j}$ is an average stochastic kernel,

$$B^{(n)}_{ai,a'j} \overset{\text{def}}{=} \frac{1}{|\Gamma^{(n)}_{ai,a'j}|}\sum_{\gamma\in\Gamma^{(n)}_{ai,a'j}}\prod_{(x,y)\in\gamma} B^{(xy)}. \tag{B.1}$$

$\Gamma^{(n)}_{ai,a'j}$ represents the set of directed path of length n joining ai and $a'j$ on $L(\mathcal{G})$ and its cardinal is precisely $|\Gamma^{(n)}_{ai,a'j}| = \big(A^n\big)_{ai}^{a'j}$.

Lemma B.2. *For any* $(x_{ai}, x_{a'j}) \in \mathbb{R}^{2q}$*, such that* $\langle x_i\rangle_{b^{(i)}} = 0$ *and*

$$x_{a'j,\ell}b^{(j)}_\ell = \sum_k x_{ai,k}b^{(i)}_k\big(B^{(n)}_{ai,a'j}\big)_{k\ell},$$

the following inequality holds

$$\|x_{a'j}\|_{b^{(j)}} \le \mu_2^n\|x_{ai}\|_{b^{(i)}}.$$

Proof. Let $x^\gamma_{a'j}$ the contribution to $x_{a'j}$ corresponding to the path $\gamma\in\Gamma^{(n)}_{ai,a'j}$. Using Lemma B.1 recursively yields for each individual path

$$\|x^\gamma_{a'j}\|_{b^{(j)}} \le \mu_2^n\|x_{ai}\|_{b^{(i)}},$$

and, owing to triangle inequality,

$$\|x_{a'j}\|_{b^{(j)}} \le \frac{1}{|\Gamma^{(n)}_{ai,a'j}|}\sum_{\gamma\in\Gamma^{(n)}_{ai,a'j}}\|x^\gamma_{a'j}\|_{b^{(j)}} \le \mu_2^n\|x_{ai}\|_{b^{(i)}}.$$

■

It is now possible to conclude the proof of the theorem.

Proof of Theorem 5.3(i). (i) Let $\mathbf{v}$ and $\mathbf{v}'$ two vectors with $\mathbf{v}' = \mathbf{v}\tilde{J}^n = \mathbf{v}(\mathbb{I}-M)J^n$, ($M$ is the projector defined in Proposition 5.2) since $\tilde{J}M = 0$. Recall that the effect of $(\mathbb{I}-M)$ is to first project on a vector with zero local sum, $\sum_k\big(\mathbf{v}(\mathbb{I}-M)\big)_{ai,k} = 0$, $\forall i\in\mathbb{V}$, so we assume directly $\mathbf{v}$ of the form

$$v_{ai,k} = x_{ai,k}b^{(i)}_k, \qquad \text{with} \qquad \langle x_{ai}\rangle_{b^{(i)}} = 0.$$

As a result $\mathbf{v}' = \mathbf{v}J^n = \mathbf{v}'(\mathbb{I}-M)$ is of the same form. Let $x'_{a'j,\ell} \overset{\text{def}}{=} v'_{a'j,\ell}/b^{(j)}_\ell$. We have

$$\|x'\|_{\pi,b} \le \sum_{a'\to j}\pi_{a'j}\sum_{a\to i}\big(A^n\big)_{ai}^{a'j}\|y_{a'j}\|_{b^{(j)}}$$

with $y_{a'j,\ell}\ b^{(j)}_\ell = \sum_k x_{ai,k}b^{(i)}_k\big(B^{(n)}_{ai,a'j}\big)_{k\ell}$. From Lemma B.2 applied to $y_{a'j}$,

$$\|x'\|_{\pi,b} \le \sum_{a'\to j}\pi_{a'j}\sum_{a\to i}\big(A^n\big)_{ai}^{a'j}\mu_2^n\|x_{ai}\|_{b^{(i)}} = \lambda_1^n\mu_2^n\|x\|_{\pi,b},$$

since $\boldsymbol{\pi}$ is the right Perron vector of A. ■

References

C. Berge. *Théorie des graphes et ses applications*, volume II of *Collection Universitaire des Mathématiques.* Dunod, 2ème edition, 1967.

P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation and queues.* Springer-Verlag, 1999.

P. Diaconis and D. Strook. Geometric bounds for eigenvalues of markov chains. *Ann. Appl. Probab*, 1(1):36–61, 1991.

C. Furtlehner, J.-M. Lasgouttes, and A. Auger. Learning multiple belief propagation fixed points for real time inference. *Physica A: Statistical Mechanics and its Applications*, 389(1):149–163, 2010.

P. R. Halmos. *Finite-Dimensional Vector Space.* Springer-Velag, 1974.

D. J. Hartfiel. System behavior in quotient systems. *Applied Mathematics and Computation*, 81(1):31–48, 1997.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. *Advances in Neural Information Processing Systems*, 15, 2003.

A. Ihler, J. I. Fischer, and A. Willsky. Loopy belief propagation: Convergence and effects of message errors. *J. Mach. Learn. Res.*, 6:905–936, 2005.

F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Inf. Th.*, 47(2):498–519, 2001.

J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Trans. on Inf. Th.*, 53(12):4422–4437, 2007.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.* Morgan Kaufmann, 1988.

E. Seneta. *Non-negative matrices and Markov chains.* Springer, 2006.

S. Tatikonda and M. Jordan. Loopy belief propagation and gibbs measures. In *UAI-02*, pages 493–50, 2002.

M. J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches.* PhD thesis, MIT, Jan. 2002.

Y. Watanabe and K. Fukumizu. Graph zeta function in the bethe free energy and loopy belief propagation. In *Advances in Neural Information Processing Systems*, volume 22, pages 2017–2025, 2009.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inform. Theory.*, 51(7):2282–2312, 2005.